# 一种基于目标检测的空间场景分类框架

吴若玲

**摘要**

长期以来，空间场景分类一直是地理信息科学领域的一个突出研究领域。在过去传统方法主要依赖于基于图像特征的检索方法。然而，随着深度学习和人工智能领域的迅速发展，对复杂空间场景的高效分类日益重要。本文提出了一种新颖的框架，该框架将目标检测与知识图谱相结合，自动完成空间场景分类。首先使用目标检测技术对输入图像进行处理以识别场景中的关键实体。随后，利用包含各种空间场景、实体及其关系的知识图谱来识别空间场景分类。为了验证该框架的有效性，我们使用八个空间场景类别进行了实验。实验结果表明，得到的分类结果与真实空间类型较为一致，验证了框架的有效性，展现了空间场景分类的潜在应用价值。

**关键字:** 空间场景分类，目标检测，知识图谱

**Abstract**

Spatial scene classification has long been a prominent area of research in the field of geographic information science. In the past, traditional approaches heavily relied on retrieval methods based on image features. However, given the rapid advancements in deep learning and artificial intelligence, the efficient classification of complex spatial scenes has become increasingly crucial. This paper presents a novel framework that combines object detection with knowledge graph to automate the process of spatial scene classification. Initially, the input images undergo processing using object detection techniques to identify key entities within the scenes. Subsequently, a knowledge graph, which encompasses various spatial scenes, entities, and their relationships, is utilized to identity spatial scene catogories. To validate the effectiveness of the framework, experiments were conducted using eight spatial scene categories as an example. The results demonstrated a high level of consistency with actual spatial types, thus affirming the efficacy of the framework and highlighting its potential application value in the domain of spatial scene classification.

**Key words:** Spatial Scene Classification, Object Detection, Knowledge Graph

## 1 引言

近年来，深度学习在图像处理和识别方面表现出了非凡的能力, 目标检测是计算机视觉领域的一个焦点，取得了显著进展。这项技术可以高效地处理复杂场景、进行多目标检测。在现有的常见的目标检测任务中，例如行人检测、交通信号灯检测、遥感目标检测等 [1]，大多是针对前景目标进行检测，而图片中的背景元素常常能够提供重要信息，帮助我们理解图像所描绘的场景，因此，检测图像中的背景元素并判断它们所属的空间场景成为了一个重要挑战。

图像背景元素检测可以帮助分析人员识别出背景中的关键物体、位置和特征，从而构建一个更准确的空间场景再现。例如，在犯罪调查中，图像背景元素检测可以帮助执法机构识别犯罪嫌疑人的所处位置和行踪，通过识别出关键的背景元素如建筑物、道路和地标，帮助警方锁定嫌疑人位置。这一技术不仅有助于法律执法机构更好地理解犯罪现场的情况，还能提供重要的证据支持，有助于犯罪调查和司法程序的顺利进行，进而有望提高犯罪调查的效率和准确性，为社会安全和正义做出更大的贡献。

基于以上的研究背景和需求，本文提出一个识别和判断空间场景类别的框架。基于深度学习的目标检测模型可以帮助我们实现图像背景元素的识别，此外，构建了空间场景知识图谱，将空间场景与场景对应的关键物体相对应。基于识别到的实体在知识图谱中进行搜索，最终获得该图像的空间场景类别，实现对空间场景的识别与分类。

# 2 相关工作

## 2.1 空间场景识别与分类

空间场景的识别和分类可以分为基于特征的图像检索方法和基于深度学习的图像检索方法 [2]。

传统的基于图像特征的检索方法是一种词袋模型（Bag-of-Visual-Words, BoVW），首先使用局部特征描述符，例如 SIFT[3] 来进行局部特征表述，再计算图像特征与视觉词典中每个视觉单词的距离，生成视觉单词直方图，进而实现图像检索并实现场景识别。

尽管传统的特征描述方法能够满足场景识别需求，但深度学习的出现，给场景分类与识别以及整个计算机视觉领域带来了突破性的性能提升。以 AlexNet[4]，VGG[5] 和 ResNet[6] 为代表的卷积神经网络为图像分类领域带来巨大变革，ImageNet[7]，COCO[8] 和 Places[9] 等大规模图像数据集的可用性进一步丰富了这一领域。场景识别更多关注于空间场景中的地理特征，借助卷积神经网络进行特征提取将助力这一研究。CNN 能够自动从图像中学习和捕获不同层次的特征，包括边缘、纹理、形状和高级语义特征，这对于空间场景识别至关重要。它能够处理不同尺度的输入图像，能够自适应地捕捉大尺度的全局特征和小尺度的局部特征，从而更好地理解复杂的场景。此外，通过数据增强技术可以扩展训练数据，提高模型的泛化能力。这对于处理不同时间、天气和光照条件下的场景非常重要。

## 2.2 目标检测

由于空间场景包含前景和背景的复杂性特点，想通过提取背景实现对空间场景的分类具有一定挑战，因此，对场景图像先采用目标检测的方法提取场景中的实体目标，在根据提取到的实体对空间场景进行分类，可以获得更好的效果。目标检测领域的发展主要分为两个主要时期：传统目标检测算法时期和基于深度学习的目标检测算法时期 [10]。

传统的目标检测方法主要有三个步骤：选择候选目标区域、提取特征和利用分类器分类 [11]。传统目标检测算法的代表有 P. Viola 和 M. Jones 等学者在 2001 年提出的 VJ 检测器 [12]、N. Dalal 和 B. Triggs 在 2005 年提出的方向梯度直方图检测器 [13]、P. Felzenszwalb 在 2010 年提出的可变性组件模型（Deformable Part-based Model，DPM)[14] 等。

由于卷积神经网络（Convolutional Neural Networks，CNN）提取特征相较人工提取更加丰富全面，因而基于深度学习的目标检测快速发展起来。通常情况下，以检测器设计结构不同，基于

深度学习的目标检测又分为了“两阶段检测”和“一阶段检测”两种类型。2014 年，R. Girshick 等学者提出了 R-CNN，开创了“两阶段检测”的先河。“两阶段检测”采用“候选框 + 预测”的模式，具有较好的检测精度。在之后的几年里，学者们相继提出了 Fast R-CNN[15]、Faster R-CNN[16]、Mask R-CNN[17]、特征金字塔网络（Feature Pyramid Networks，FPN）[18] 等。针对两阶段目标检测算法的低效问题，在 2016 年，R. Joseph 等学者提出 YOLO（You Only Look Once）模型 [19]，这是基于深度学习的首个一阶段目标检测器。YOLO 系列模型仅设计了一个卷积神经网络，输入整张图像，然后把图像分为多个区域，对每个区域分别预测出边界框和类别置信度。虽然在准确度上稍逊色于“双阶段模型”，但胜在速度较快，更加轻量级，同样具有较好的应用前景。YOLO 系列模型一直在被改进和优化 [20, 21, 22]，在 2023 年 1 月，最新的 YOLOv8 问世，学者们对其主干网络的模块以及损失函数进行优化，具有较好的效果。此外，近些年来也涌现出很多单阶段模型：Single Shot MultiBox Detector(SSD)[23]、RetinaNet[24]、EfficientDet[25] 等。

### 2.3 知识图谱

知识图谱是人工智能和自然语言处理领域的一个重要概念，它是一种将知识以图谱形式组织和表示的方法。知识图谱的概念可以追溯到上世纪六十年代的“语义网络”，它采用相互连接的节点和边来表示知识，节点表示对象、概念，边表示节点之间的关系，已经具备如今知识图谱的雏形。2012 年，谷歌提出“知识图谱”的概念，它最早是应用于改善搜索引擎的搜索结果，其目标是让用户不必导航到其他网站并自己汇总信息，即可通过知识图谱提供的结构化信息来解决他们查询的问题。在如今的信息化时代，知识图谱不仅用于搜索引擎改进，还在自然语言处理、推荐系统、智能助手、医疗保健、智能交通等领域得到广泛应用。它们帮助机器理解上下文、进行推理和提供智能决策支持，推动了人工智能领域的发展。随着开放数据的增加和开源工具的发展，知识图谱的创建也变得更加容易。一些大型知识图谱项目，如 Wikidata 和 DBpedia，开放了大量结构化数据，供研究人员和开发者使用。

知识图谱是一种用于组织和表示知识的图形数据结构，它由实体（Entities）和它们之间的关系（Relationships）组成。实体（Entities）是知识图谱中的核心元素，通常表示现实世界中的事物、概念、个体或对象，它可以是具体的物体，如人、地点、书籍，也可以是抽象的概念，如情感、关系、事件等。属性（Attributes）是与实体相关的属性或特征，用于描述实体的各种方面，有助于详细描述实体，并提供了更多关于实体的信息。关系（Relationships）表示实体之间的连接或联系，它们描述了不同实体之间的关联性。关系通常具有名称和方向，例如，“包含”是一种关系，可以连接书店实体和书实体。关系可以是单向的或双向的，还可以具有属性来描述关系的属性，如权重、强度等。

知识图谱的结构化特性使其成为处理大规模知识的有效工具，促进了语义搜索、自然语言处理、智能推荐和决策支持等领域的发展。在我们的任务上，知识图谱可以存储空间场景和场景中的实体之间的关系，将二者进行匹配和关联，使得可以根据检测到的实体判断空间场景的类型，具有一定的实际意义。

## 3 空间场景分类框架

为了实现空间场景的分类，我们提出了一种基于目标检测和知识图谱的综合框架。该框架共分为两个部分：（1）基于目标检测的空间实体识别;（2）基于知识图谱的空间场景分类。框架如图 1 所示。

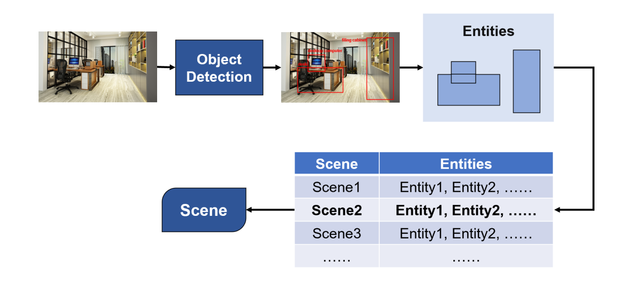



图 1: 空间场景分类方法框架

Figure 1: The Framework of Spatial Scene Classification

表 1: 空间场景与其实体的对应关系

Table 1: Correspondence between Spatial Scenes and Entities

| Scenes | Entities |
|---|---|
| office | desk, desktop computer, notebook computer, filing cabinet, folding chair, printer |
| kitchen | frying pan, pitcher, plate rack, refrigerator |
| restaurant | dining table, menu, plate, hot dog, cabbage |
| city | check, traffic sign, traffic light, bookstore, shoe store |
| coast | seashore, shoal, yawl, swimsuit |
| mountain | alp, volcano, cliff, valley |
| forest | hay, cardoon |
| countryside | barn, boathouse, thatched roof, ox, hen, duck, goose |

### 3.1 空间实体识别

在我们的空间场景分类框架中,首要步骤是使用目标检测技术来处理输入的图像或视频。You Only Look Once(YOLO) 模型以其轻量和易于训练的特点,如今被广泛使用。对于这项任务,YOLO 模型可以很好胜任，所以选用此模型。这一阶段旨在从图像中识别和定位出现在场景中的不同物体或实体，确定它们的位置和类别。

### 3.2 空间场景分类

(1) 知识图谱构建

目前存在大量的空间场景数据集，有些数据集以室内场景为主，如 MIT67[26] 和 Interior-Net[27]，室外场景数据集则更多地面向自动驾驶领域，例如 Cityscapes[28]。此外，还有很多数据集同时包括室内和室外场景，例如 SUN397[29] 和 Places[9] 等。这些公共数据集有效地满足了空间场景分类和各种其他应用等任务的要求。在上述公共数据集的指导下，在此阶段中构建空间场景及其空间实体的知识图谱。空间场景和空间场景中的实体无法穷尽，因此，只列举了一组常见的空间场景并选取其最关键的元素，将其对应匹配。具体关系如表 1 所示。

(2) 知识图谱查询

利用前一阶段的目标检测结果作为查询条件，可以与构建好的知识图谱进行交互。知识图谱包含了不同空间场景、实体以及它们之间的关系，通过匹配目标检测结果中的实体和知识图谱中

的实体，推断出当前图像对应的空间场景类别。例如，如果检测到的实体包含 desk、filing cabinet 和 printer，并且知识图谱中定义了这些实体的包含关系，那么我们可以确定当前场景为 office。

# 4 实验

## 4.1 数据集

数据集的数据量与图片清晰度决定了模型的拟合能力和预测泛化能力，因此，数据集的选取至关重要。ImageNet 数据集是图像分类、检测、定位的最常用数据集之一，其多样性和广泛性使得它成为图像理解和场景识别研究的重要资源，和提出的框架的任务比较适配。该数据集中包含大量带有类别和位置标注信息的图片，故在目标检测训练阶段，从 ImageNet 数据集中选取和任务相关的实体进行实验。选取的数据包含了 37 个不同类别的图像，各类别图像数量较为均衡，共计 22019 张图片，包含 25630 个目标实体。

## 4.2 实验设置

YOLOv5 模型训练在 Ubuntu 18.04 系统下使用 GPU 进行加速，使用搭载 GA102 GPU 的 GeForce RTX 3080 显卡。模型训练的参数设置：batch 大小为 16，学习率为 0.01，迭代次数为 300 代。

空间场景知识图谱的构建基于高性能的 NoSQL 图形数据库 Neo4j，并实现可视化展示。使用 python 实现和 Neo4j 数据集的连接，并对数据库进行查询。

## 4.3 实验流程

为验证提出的框架的有效性，选择了 8 个具有代表性的空间场景进行了实验。这些场景都具有较为标志的目标实体，场景包括 3 个室内场景：office、restaurant 和 kitchen，5 个室外场景：city、forest、coast、countryside 和 mountain。从网络上收集了 240 张测试图片，它们属于这些不同场景，每一类别平均 30 张，采用 YOLOv5 模型进行目标检测，产生了每个图像中检测到的物体的位置坐标和类别标签。

根据目标检测的结果查询知识图谱，建立检测到的实体与知识图谱中的实体之间的联系。将检测到的实体依次对应到知识图谱中，并对查询到的空间场景进行计数，如果存在最高数量，则判定为该场景，若计数最高数量有多个，则结合检测到的目标实体的置信度共同判断，得到预测场景。由此，基于知识图谱查询的结果，可以确定每个图像的预测空间场景类别。

# 5 结果

混淆矩阵是用于评估分类模型性能的一种矩阵形式的工具，它可以提供更全面的性能评估，避免仅仅依赖准确率等指标导致误导。本实验以混淆矩阵的形式呈现分类结果，如图 2 所示。

从实验结果来看，对于所选的 8 个场景，大部分场景的图片被归到正确的类别，显示出其在不同场景上的有效性和通用性，少部分图像由于在目标检测阶段未识别到目标实体，影响了对后面场景分类的判断。部分场景之间具有一定的相似性，比如 office 和 restaurant 都可能存在 table，mountain 和 forest 都可能存在 valley 等等，在检测与分类任务上存在小误差。总体来看，本文提出的框架在空间场景分类任务上表现良好。具体实施结果如表 2 所示。

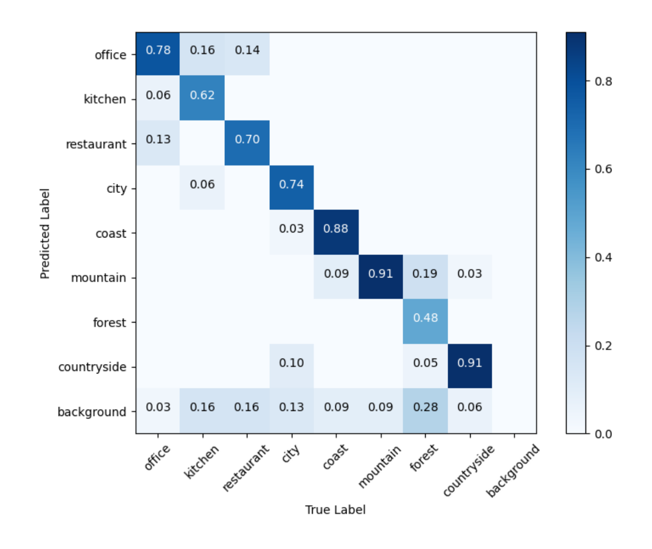



图 2: 空间场景分类结果

Figure 2: Spatial Scene Classification Results

# 6 结论与展望

本文提出了一种基于目标检测和知识图谱的空间场景分类框架，旨在实现对图像中的场景进行分类和理解，通过 8 个场景的测试，验证了框架的有效性。通过融合计算机视觉领域的目标检测技术和知识图谱的语义表示，成功地将视觉感知与结构化知识相结合，为空间场景分类任务带来了新的可能性。

在应用层面，在司法和公共安全领域，应用此框架来识别犯罪场景具有广阔发展前景。它可以协助进行案件调查，对犯罪现场进行分析与证据收集，进而对重建犯罪场景，这为法律执法提供了强大的技术支持，有助于提高司法系统的效率，促进公共安全，增加社会安全感。

本文研究仍有局限性，在两个方面有潜在的改进空间：（1）扩大场景的种类和知识图谱的规模一定可以让框架更加通用，覆盖更多实际情景。(2) 将更先进的决策方法集成到从知识图查询到最终分类结果的过程中可以增强框架的可解释性。

# 参考文献

[1] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023.

[2] 王亚军, 穆杞梓, 余末银, 王洪海, 朱立远. 基于卷积神经网络的室外场景识别. *自动化应用*, 64(14):201–207, 2023.

[3] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84–90, 2012.

[5] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, May 2015.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[8] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

[9] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.

[10] 许德刚, 王露, 李凡. 深度学习的典型目标检测算法研究综述. *计算机工程与应用*, 57(8):10–25, 2021.

[11] 张顺, 龚怡宏, 王进军. 深度卷积神经网络的发展及其在计算机视觉领域的应用. *计算机学报*, 42(3):453–482, 2019.

[12] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001.

[13] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005.

[14] Pedro F. Felzenszwalb, Ross B. Girshick, and David McAllester. Cascade object detection with deformable part models. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2241–2248, 2010.

[15] Ross Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[18] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017.

[19] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[20] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017.

[21] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *ArXiv*, abs/1804.02767, 2018.

[22] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *ArXiv*, abs/2004.10934, 2020.

[23] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision*, pages 21–37, 2016.

[24] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis  Machine Intelligence*, PP(99):2999–3007, 2017.

[25] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10778–10787, 2020.

[26] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–420, 2009.

[27] Wenbin Li, Sajad Saeedi, John McCormac, Ronald Clark, Dimos Tzoumanikas, Qing Ye, Yuzhong Huang, Rui Tang, and Stefan Leutenegger. Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. In *British Machine Vision Conference (BMVC)*, 2018.

[28] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[29] Luis Herranz, Shuqiang Jiang, and Xiangyang Li. Scene recognition with cnns: Objects, scales and dataset bias. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 571–579, 2016.

表 2: 空间场景分类框架分类效果

Table 2: Spatial Scene Classification Framework Classification Effects

| Scene Images | Detected Entities | Knowledge Graph Queries | Classification Results |
|---|---|---|---|
| | | | office |
| | | | kitchen (after comparing confidence) |
| | | | restaurant |
| | | | city |
| | | | coast |
| | | | mountain |
| | | | forest |
| | | | countryside |